Methodological Insights from a Rigorous Small Scale Design Experiment

Abstract
This paper discusses the methods used to conduct high quality small-scale design experiments. It aims to provide a demonstrator promoting the uptake of more rigorous methods in design research and based on this it aims to specify a body of further work for linking study types and contexts. A small-scale experiment was conducted using methods specifically developed to mitigate four core problem areas identified from review: context, system understanding, methods and controls. The techniques were then critiqued in detail and used to draw several insights for design researchers including the value of control techniques and triangulation of metrics. Finally, the critique is used to specify further research aimed at linking design experimentation and design practice more effectively for design research.

Keywords: Research methods, design research, design science, case study, experiment

Design Studies
Authors: Philip Cash*, Edward Elias, Elies Dekoninck, Steve Culley

*Corresponding author

Affiliations:

University of Bath, UK

Design researchers have used experiments and observational studies extensively over the last forty years to explore the working practises and performance of designers and design teams (Cross, 2007). Recent examples include Howard et al.'s (2010) work on ideation, Dong's (2005) work on analysing design team communication, Bakeman & Deckner's (2003) work on behaviour others across a range of areas (Ball & Ormerod, 2000; M. A. Robinson, et al., 2005). Empirical study forms a valuable part of design research, providing essential insight into many areas of design whilst also supporting theory-building (Stempfle & Badke-schaub, 2002) and the development of real world impact as emphasised by Briggs (2006) and Cross et al. (1996) . However, there is an ongoing challenge to improve the quality of empirical studies in design research (Blessing & Chakrabarti, 2009).

One standard approach to improving quality has been to develop large-scale statistical studies, however, these are very time/resource intensive and not always appropriate for design research topics as such are rarely used in design research. A common technique that is widely used is small-scale studies. Thus, this paper examines how small-scale studies can be made more rigorous, with the aim of providing a demonstrator to support uptake of underutilised methods in design research. To this end, a laboratory-based design experiment has been developed to explore what methods can be used to improve validity, replicability and reliability. Although these studies are not a substitute for large-scale statistical validation, this paper will show that, with improved rigorous methods, small-scale studies can show possible trends and give insights into design situations.

This is demonstrated using a case study experiment with the hypothesis: *design teams benefit from having design relevant information presented to them during the early design phases of a product development process*. The experiment also aimed to investigate what format would be most effective for the pushed information: video footage of users interacting with the product or numerical data describing the same interactions. Five design teams were each tasked with generating design ideas for a domestic refrigerator with the aim of reducing the amount of electrical energy wasted by the user through poor or inefficient use. Improvements in the energy efficiency of the users' behaviour were specified by the researchers to be achieved through the physical design of the product and not by improving the energy awareness or education of the user. In order to effectively tackle this task, designers require knowledge of user behaviour: how they use refrigerators, the rationale for their actions, and where/when inefficient use occurs. These behaviours were collected and discussed by Elias et al. (2009) and have been used to inform the various types of information stimuli presented to the teams. A more detailed analysis of the primary experimental results will be presented elsewhere.

This study was selected as it represented a typical small-scale design research study (see Section 1). This paper does not, however, focus on the hypothesis-related results of the experiment; instead, it gives a critique of the experimental methods, their affect on the study, and identifies a number of

methodological lessons. This critique is then used to discuss the role of small-scale studies in design research and the future work needed to support it. The first necessary step in critically appraising an experimental method was the identification of major methodological problems likely to be encountered. These problems were identified based on a literature review of design research and its contributing fields (Friedman, 2003). The problems synthesised from this review provided a basis for the identification and development of the mitigation approaches used in this study such as the placebo control group. These problem areas also formed the basis for a critical appraisal of the experimental methods typically used in design research and, subsequently, a specification of further work that demands a distinct and significant body of work beyond the scope of this paper.

1 Experimental problems

Throughout design research there has been a drive to improve the quality of empirical research, identifying validity and reliability as critical success factors for quality, uptake and impact (Blessing & Chakrabarti, 2009; Dillon, 2006; Dyba & Dingsoyr, 2008; Lanubile, 1997; M. A. Robinson, 2010; Sharp & Robinson, 2008; Valkenburg & Kleinsmann, 2009). Although the specific circumstances in which problems are encountered vary, there is much commonality in the form and scope of the overarching problems (Cash, et al., 2009; Friedman, 2003). Drawing on the literature from contributing fields (disciplines outside design identified by Friedman (2003)) and design research, shows that there are numerous appropriate mitigation techniques. Mitigation, in this context, means the reduction or elimination of problems affecting validity, replicability and reliability with respect to design research experimentation. Consequently some of these techniques have been implemented in this study to present a more rigorous small-scale design experiment.

The problem areas collectively affect all types of validity (Adelman, 1991), impact and ultimately, uptake (Glasgow & Emmons, 2007). Exemplar reviews from the contributing fields, emphasising issues associated with lack of experimental planning and effective mitigation techniques, are: Glasgow (2007) in clinical research, Gorard and Cook (2007) in education research, Adelman (1991) in decision support systems research and Blessing and Chakrabarti (2009) in design research. Hansen et al. (2001), from design research, highlight complexity and the need for common methodological approaches. The importance of capturing the environmental conditions and the context in which a participant is acting is also highlighted. Dillon (2006) emphasises that the mind works in a dynamic interaction with the environment and the context of the task. Thus, it is important to contextualise both the task (Lave, 1988; H. Robinson, et al., 2007) and the research (Sharp & Robinson, 2008) as well as reporting factors such as methods, environment and population – all of which are key issues within design research.

In light of these issues it is important to clarify the value and role of this paper; three arguments are presented. Firstly, seven of the most recent small-scale experiments in design research journals were

specifically reviewed (Table 1). Although each of these studies makes significant effort to address methodological issues, several problems are evident in each. Therefore, the methodological problems summarised below are significant and very relevant to current design research – particularly the use of control procedures. Secondly, small-scale scoping studies, although not always ideal, play an important role in design research for pragmatic as well as methodological reasons. Thirdly, despite the relevance of some mitigation techniques in other fields there is limited uptake in design research. Table 1 takes four common issues/techniques and rates each paper either 'ok' or '-' (indicating failure to implement the technique effectively).

Study	**Summary**	**Relevant Issues**			
		No-treatment control	Placebo control	Discussion of limitations	Population/ methods description
(Corremans, 2009)	A pre and post-test study using students to assess a design method	-	-	-	Ok
(Kurtoglu, et al., 2009)	A small study using students to assess the value of a computational approach	-	-	Ok	Ok
(Cai, et al., 2010)	A small experiment looking at sources of inspiration using multiple short tests	-	-	-	-
(Stones & Cassidy, 2010)	A small experiment using students to assess 2 different mediums for reflection	-	-	-	-
(Lemons, et al., 2010)	A small study using students to assess the benefits of model building in teaching	-	-	-	-
(Collado-Ruiz & Ostad-Ahmad-Ghorabi, 2010)	A small study using students to assess the effect of information on creativity	Ok	-	-	-
(Lopez-Mesa, et al., 2009)	A small study using students to asses the affects of stimuli on idea finding	-	-	-	Ok

Table 1: Recent small-scale empirical studies in design research – a brief examination

Based on the review of design research as well as the specific review outlined in Table 1 the problem areas can be categorised into: lack of contextualisation, insufficient system understanding, idiosyncratic Method implementation and insufficient control and normalisation (Table 2). Two additional underlying problems were identified as a lack of theory building and critical review. These, however, are not discussed here as they fall outside the scope of experimental methods, instead the review process, editorial boards and the community must support these. In summary a number of interlinked problems were found to contribute to issues of validity, reliability and replicability, compounded by the lack of necessary details on research methods, data collection, context and data analysis (Bender, et al., 2002).

Problem	Problem Definition
Context	The failure to adequately define or record context – this includes social context, cultural context and the context of the activity. This can lead to problems associated with maturation or the background of the subject being insufficiently accounted for. In a technical sense this can be the failure to record methods, environment or population.
System Understanding	The failure to fully explore, characterise and report the underlying variables and mechanisms at work in a test system. This negatively effects implementation of control techniques as well as affecting applicability.
Method Implementation	The inadequate definition of methods and terms, the lack of standardisation and the lack of consistency in experimental planning, recording and reporting especially with regard control and normalisation techniques.
Control and Normalisation	The inappropriate or insufficient use of control and normalisation techniques such as placebos, no-treatment control teams and deviant case analysis to give baselines for comparison. This can lead to false assurance or disproportionate results.

Table 2. Experiment problems typical of design research

From these issues, two conclusions have been drawn – forming the core mission statements for this paper:

1. In order for new techniques and approaches to gain greater acceptance in design research there needs to be clear and rigorous demonstrators showing how new methods can be applied and the rewards they offer – Achieving this for small-scale studies is the primary goal of this paper.
2. In order to address issues of external validity and reliability of small-scale studies in general there is a need for a significant body of work to be undertaken – Clarifying this need and specifying the form of this work is the secondary goal of this paper.

In order to tackle these mission statements the experiment reported here, has addressed these problems using interlinked techniques. Primarily, a placebo control approach was developed using both no-treatment and placebo control teams in an attempt to control and normalise for experimental effects as well as to provide information on underlying variables (Adair, et al., 1990). In addition, emphasis was placed on control and contextualisation (Dillon, 2006) wherever possible based on works such as Grey and Salzman (1998), Kitchenham et al. (2002) and others (Blessing & Chakrabarti, 2009; Dyba & Dingsoyr, 2008). Torgerson (2003) and others (Adair, et al., 1990; Leber, 2000) highlight placebo controls, in particular, for empirical trials. A detailed method for the placebo was developed but is summarised for brevity. In addition triangulation of metrics and analysis methods was also used as seen in the works of Cross et al. (1996) and D'Astous et al. (2001). This employed both qualitative and quantitative analysis, as discussed by Onwuegbuzie & Leech (2006), of the experimental data, allowing differences between measures to be identified and discussed in detail, improving the depth of

understanding possible. Sections 2 and 3 detail these experimental methods and are used to give context to the discussion of the mitigation methods in Section 4.

2 Experimental setup

The experiment consisted of five teams each made up of three participants. All the teams were given the same initial information:

- A handout with background information on designing products to reduce the energy impact of poor use.
- A brief describing the participant's role as designers and asking them to design a new domestic refrigerator that reduced the energy impact of poor use.
- Background information and some details on the product's target audience.
- A session plan detailing how long the teams were allowed for the task and what they would be required to produce at the end.

The brief outlined the participant's role as designers working for a new company, unconstrained by an existing design portfolio. The brief also specified that ideas should be feasible with current technology. Care was taken to focus the brief on user behaviour and the use of the refrigerator without specifying what the critical behaviour(s) might be. Precautions were also taken to avoid giving technical examples or existing solutions, instead the background documents focused on the need and the user. The documents were designed in this way to prevent the participants fixating on a specific design type while still giving them enough information to focus on the desired problem. This approach of manipulating fixation effects – with a problem rather than existing design focus – was heavily influenced by the discussion of fixation presented by Cross (2001).

The experiment was facilitated by and presided over by a researcher referred to as the 'experiment controller'. The role of the experiment controller was to present the briefing documents for the participants to read and manage the progress of the experiment, introducing new information as specified. Direct comparisons in the setup and delivery of this experiment can be made with the XeroxPARC Design Activity Workshop in Cross et al.'s (1996) design experiments. There are, however, several key differences. Firstly, in the experiment reported here, all design activity was focused on a single central table with no physical artefacts for the participants to interact with. This was done to make monitoring the design activity easier to manage and review. Secondly, the experiment controller was not allowed to respond to participant questions. This reduced possible variation between the teams as well as possible experimenter effects. The target audience, for the new product, was specified as a prototypical, young, physically able family – selected to avoid niche designing and to aid in idea comparison. The teams were given two hours for the experimental task and were asked to:

develop as many ideas as possible, evaluate those ideas, and present three final concepts. A more detailed discussion of the experimental timeline is given in Section 2.2. Finally, Table 3 and Figure 1 detail the capture equipment used.

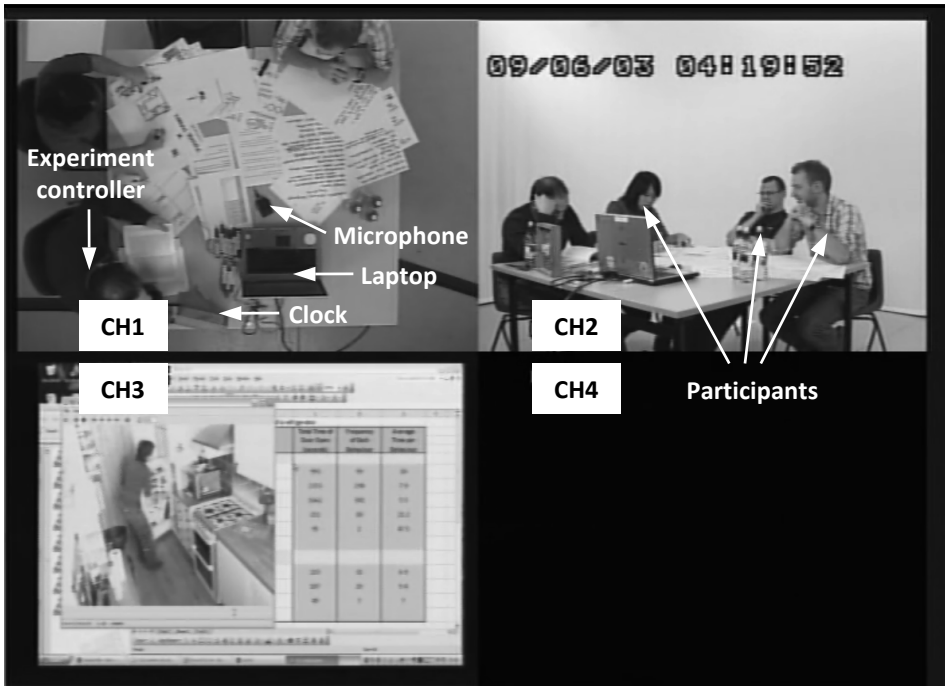

Figure 1. An example screenshot of the experimental video feed in action (CH denotes channel as explained in Table 3)

Equipment	**Description**
Video	2 cameras (CH1 from above and CH2 from the side, Figure 1)
Audio	1 microphone (in the centre of the Table)
Notes and Drawings	A3 Paper and 4 different colour pens (changed at the start of each experimental phase, see section 5) all collected at the end of each separate stage
Computer	Laptop screen feed (CH3, Figure 1) showing the information being accessed by the participants

Table 3. Breakdown of experimental equipment

The key metrics for the experimental hypothesis (*design teams benefit from having design relevant information presented to them during the early design phases of a product development process)* are outlined below with the success criteria noted in *italics* with several of these metrics being drawn from a similar study by Shah et al. (2003):

- The total number of ideas generated – *an increase in total number*.
- The originality of the ideas generated – *an increased variety and originality of ideas*.

- The effectiveness of the ideas with respect to the brief – *an increase in the number of effective ideas and a reduction in the number of irrelevant ideas generated*.

The hypothesis was tested by comparing the outputs from the teams with different types of additional information against a baseline produced by the placebo and no-treatment teams. All the participants were selected from a relatively homogeneous population of postgraduate researchers with the major contrived difference between the teams being the additional information pushed to them. Obviously there is a high degree of interpersonal variability in any population and thus a large portion of the method was spent contextualising and attempting to mitigate these differences (Sections 2.1 and 2.2). In other words the composition of the teams (Bell, 2007) can have a large effect on the results, thus team formation is detailed in the next section.

2.1 Team formation

The participants were selected from a population of postgraduate students at the Department of Mechanical Engineering, University of Bath. This was carried out randomly to reduce variables such as, levels of creativity, work rate, team cohesion and others (Ancona & Caldwell, 1992). All participants had received academic training to at least masters degree level and had experience in academic engineering design research, the engineering design process, brainstorming and the general creativity processes. They also had similar academic backgrounds (through the British university system), professional backgrounds (an average of 3 years working as post graduate students) and professional focus (working in the same broad research area). Due to the limited number of participants (15), teams were formed and balanced using Belbin Team Role scores while maintaining the highest level of randomisation as advocated by Torgerson & Torgerson (2003). This was achieved by anonymzing and randomising the participants' prior to the Belbin testing, ensuring that bias affecting the final selection of the teams is limited. Finally, teams were randomly assigned the experimental conditions: additional information 1, 2, 3, no-treatment control and placebo control.

2.1.1 Team size

A second key consideration was team size, highlighted by several authors (Brewer & Kramer, 1986; Drach-Zahavy & Anit, 2001; Stewart, 2006). Opinion on optimal team size varies with some studies showing that larger teams produce more ideas (Campion, 1993; Guzzo & Dickson, 1996; Hare, 1952) while others dispute this (Hackman & Vidmar, 1970; Hwang & Guynes, 1994). In general larger teams tend to take longer to reach a decision and require clear leadership to be consistently effective. This is due to the fact that member dissatisfaction increases and participation/contribution decreases with size (Cummings, et al., 1974; Gorla & Lam, 2004). However, small teams show higher levels of tension and what Hoffman (1965) calls "ideational conflict", preventing them from quickly settling on a single idea. This conflict makes them more conducive to creative problem solving. In summary, balancing the

conflicting opinions on optimum group size gives a team size of between three and five, depending on task (Table 4).

Team Size	Participants Needed	Recording Method	Drawbacks / Benefits
1	5	Concurrent Verbalisation	A single strong / weak participant may affect results. Not a suitable representation of industrial teams that are normally three or more people in this situation.
2	10	Listen to Discussion	A single strong / weak participant may affect results, but two people removes the need for verbalisation as their discussion can be recorded easily.
3	15	Listen to Discussion	Strong / weak participants are balanced amongst other team members. Participant discussion is easy to follow. No parallel discussions possible.
4	20	Listen to Multiple Discussions	Strong / weak participants are balanced. Greater idea generation potential. Multiple parallel discussions may be hard to follow. Lots of people required.
5	25	Listen to Multiple Discussions	The same drawbacks and benefits as having 4 people per team but the literature suggests they would also require formal team leadership to be most effective.

Table 4. Team size drawbacks and benefits matrix, with chosen size of three highlighted

In addition, there were logistical requirements to try and record the discussions and actions of the team. As team size increased the difficulty in recording these different aspects also increased. However, small teams (one or two people) increase the amount of silent 'thinking' time where audio and video recording are less effective. Recording small teams relies on 'thinking aloud' protocols of concurrent verbalisation where a participant gives a continuous narration of their thoughts. Although these types of protocol can be effective there is debate as to the level of effect they have on the participants design process (Cross, et al., 1996; Gero & Tang, 2001). The other major drawback to using small teams is that they are not representative of the industrial situation where teams are larger with significant differences in the behaviours of individuals and dyads when compared to larger groups (Hackman & Vidmar, 1970; Salas, et al., 2008). Alternatively, there are no significant differences between groups of three, four or five (Baltes, et al., 2002). Selecting a team size of three addresses many of these issues and provides a number of experimental benefits including: the ability to balance the teams by spreading participants of varying Belbin score, avoiding participant alienation, eliminating the need for concurrent verbalisation and eliminating the possibility of parallel conversations – simplifying transcription and analysis.

2.1.2 Team balancing

With a team size of three, it was important to balance the teams effectively in order to limit performance variability. Team balancing was based on Belbin Team Role tests (Table 5). Belbin Team Roles are one of the most widely used assessment frameworks for measuring people's character (Senior, 1997) and are commonly used in interview situations. Belbin roles were chosen over another popular framework, the Myers-Briggs Type Indicator, because Belbin's character classifications were more logically connected with the experimental task. Belbin originally described eight possible Team Roles (Henry & Stevens, 1999) that were later expanded to nine. Team role is defined as "a tendency to behave, contribute and interrelate with others in a particular way" and are assessed using a series of questions set out by the Belbin website (Belbin, 2010).

The majority of the participants showed a spread of points over several Team Roles (Table 5), which is common (Belbin, 2010). The authors considered the "innovator" and "shaper" roles as most significant for the experimental task. Each team was thus balanced primarily on the basis of these two roles. The "innovator" role is creative, an 'ideas person' and problem solver. The "shaper" role is more dominant, a task focused leader, who will guide others towards achieving specific aims. This was selected as the secondary criteria for two reasons: firstly, they would help ensure the teams stuck to the task, and met the demanding deadlines of the study; secondly, it more realistically reflected the work environment where there is often a more senior meeting organiser/leader driving the team towards objectives.

Each team was allocated a strong "innovator", with a score above 10 points, and secondary to that condition a strong "shaper", again with a score above 10 points. All other scores were balanced as much as possible given pragmatic considerations such as participant availability for experimental time slots. In addition, friends or working colleagues – participants with shared working experience, working space or close relationships, were separated and spread amongst the other teams randomly. These relationships were assessed based on discussions with the participants and the researchers own knowledge of the participants' working relationships.

Team	Person	Belbin Team Roles								
		Coordinator	Shaper	Innovator	Evaluator	Implementer	Team Player	Networker	Finisher	Specialist
1	Person A	7	8	11	3	11	4	24	0	2
	Person B	9	28	8	2	15	2	5	0	1
	Person C	5	13	6	8	13	6	0	17	2
2	Person D	5	9	12	8	13	6	5	0	12
	Person E	2	15	7	9	13	3	9	7	5
	Person F	4	5	1	6	12	11	4	11	14
3	Person G	13	10	15	2	6	6	13	5	0
	Person H	9	2	2	7	9	11	8	3	19
	Person I	1	6	7	8	12	10	4	14	8
4	Person J	6	7	14	4	10	5	16	1	7
	Person K	0	0	6	17	19	14	0	3	11
	Person L	1	14	4	13	5	6	0	17	10
5	Person M	5	12	12	4	4	6	20	4	3
	Person N	0	3	7	18	8	8	4	13	9
	Person O	4	12	8	11	6	6	2	11	10

Table 5. Belbin Team Roles results for the five teams

2.2 Experimental procedure

The aim of the experimental procedure was to give each team comparable activities while still allowing for the different test conditions. The teams were given two main tasks:

- To generate ideas for features or products to reduce the energy loss from inefficient or wasteful user behaviour.
- To combine and/or select from these ideas the three "most effective and feasible" designs, as outlined in the brief.

Additional information was provided to four of the five teams after twenty minutes to act as creative stimuli. Three of the information sets detailed inefficient user behaviour to help focus the team on the brief and provide relevant help with the experimental task. Each of these three sets provided the same behaviour information in a different format (Table 6). The fourth information set was a placebo intervention that gave task-neutral information. The data was drawn from the same situations as the videos but obviously emphasised different aspects – e.g. the video showed the people's actions while

the data gave information such as door opening times. Although the information in both formats emphasises different things video alone is the industry standard and as such part of the experimental aim was to explore the role of additional or alternative information formats.

Team	Title	Description	Comment
Team 1	"Control"	No information, no treatment control team	No additional information provided. No interruption.
Team 2	"Placebo"	15 minute task neutral video, placebo control team	A 15 minute video of two people discussing their kitchens, the appliances they had and general appearance.
Team 3	"Video"	15 minute active video, treatment team	A 15 minute film of the refrigerator being used, including discussion and actions showing details of how and how often it is used.
Team 4	"Data"	Data pages, treatment team	A paper list of different interactions with the refrigerator and their actual energy impacts including real data on which foods/drinks were most commonly taken out of or put into the refrigerator.
Team 5	"Data + Clips"	Data pages and videos, treatment team	Same data as team 4 and a series of eight silent, hidden camera video clips demonstrating each of the behaviours, totalling approximately 13 minutes of footage.

Table 6. Team setup and additional information

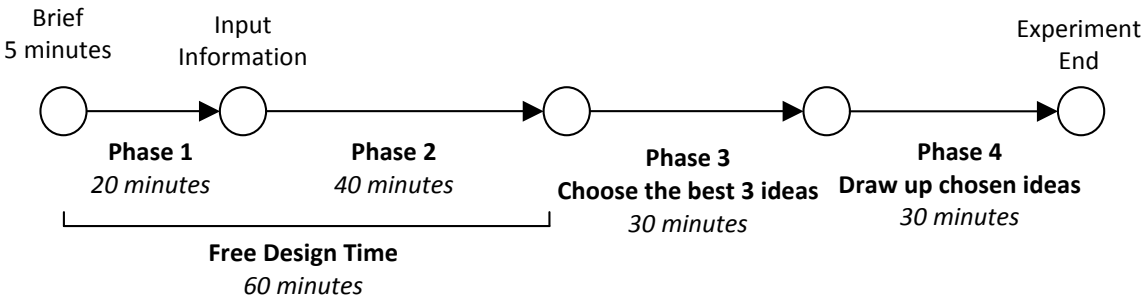

Figure 2. Experimental Timeline

Prior to the study all the participants were given basic information outlining the size of the teams, the length of time involved and the level of personal preparation required (none). Prior to the experiment none of the participants were aware that additional information was to be provided to some of the teams. This prevented teams from becoming expectant of, or simply waiting for the additional information. It also allowed for a 'no treatment control' team, which would receive no additional information – forming a baseline. Thus, the major difference between the teams was the additional

information (including its format) provided to them (Table 6). The 15 minute placebo video (team 2) involved two people discussing their kitchen appliances and general appearance. The length and style of the video was similar to that of the hypothetically 'active' additional information video (team 3) but included no specific information about refrigerator use. This is known as an 'Act+' type placebo control team (Adair, et al., 1990) normalising for the disruption of introducing the additional information – forming a second baseline. The placebo video was selected by assessing potential candidates against a list of variables that could influence the participants and experiment. This list was split between the hypothetically 'active' variables and those considered 'non-active'. Potential placebos were then assessed until one was found which had what was considered by the authors to have very little effect on the 'active' variables – a more detailed breakdown of how this method was implemented has been specifically made available online – www.designresearchmethods.com.

The experiment was divided into four phases (Figure 2). Phases 1 and 2 were free design time in which the teams could complete the first task of generating ideas. There were no methods prescribed to the teams during these phases. At the end of Phase 1 four of the five teams were interrupted and given additional information (Table 6). Phase 3 gave the teams additional time to generate and develop ideas, but also to select the best three. Finally, phase 4 gave the teams time to develop a sketched explanation of the three final ideas. The total time was just over two hours. This was selected as a suitable length of time to allow the teams to go from the design brief to a finished idea while keeping the disruption to the participants to a minimum. The two hours was split between design divergence, the free design time (Phases 1 and 2), and design convergence, the idea assessment and final drawing up (Phases 3 and 4). Participants were not aware of this division until after the study.

Experiment controller's script: (boxed)

"In front of you is all the information you require for this experiment, the briefs, some A3 paper, pens and a clock. I ask you to undertake this experiment in good faith and to take on the roles of the designers, as described in the design brief. I will sit here but take no part in the experiment. I will at certain times prompt you to move on to the next stage of the experiment. At times during the experiment I will replace your pens with different coloured ones as this will help with our review process. I cannot answer questions during the experiment or help you in any way. Please start by reading the sheets in front of you. You now have a few minutes to read this information and collect your thoughts before the experiment starts. I will tell you when you can move on to the next stage."

Phase 1 After a five minute introduction, during which the experiment controller read the scripted instructions (see above – boxed) and gave the participants time to read the briefing material, which included a session plan for the experiment (see below), paper and pens were provided and the experiment proper began.

Session Plan:

This exercise will take 2 hours and 5 minutes and is divided into the following sections:

1. *You have 5 minutes to be briefed, review the information provided and collect your thoughts.*
2. *For the first hour, you are asked to brain storm and come up with as many different ideas as you can.*
3. *You will then be given 30 minutes to review your ideas and choose the 3 most effective and feasible product ideas.*
4. *You will then have another 30 minutes to develop these 3 ideas and sketch each on a separate piece of A3 paper. The ideas should be understandable from this piece of paper alone.*

Phase one was the same for all five teams, at the start of which the teams were instructed that they had 60 minutes to develop as many ideas as possible, however after 20 minutes four of the five teams began phase 2 and were given additional information. This initial period of 20 minutes was the same for all five teams and allowed a baseline comparison to be made between the teams.

"Ok, I am going to give you some pens, you now have one hour to come up with as many ideas as you can."

Phase 2 After phase one, four of the teams received additional information and continued with idea generation for a further 40 minutes.

"I have a video (or "some information", in the case of the data) for you to look at which may be of some help."

At this point a laptop was opened and the video/data displayed. Until this point the laptop had been closed on the table so as to not raise expectations of what might be happening. The laptop was present for all the teams. Once the video was finished, the controller instructed the teams that if they wished to re-watch any or all of the footage, they could at any time by using the laptop. The data was similarly made available after its initial introduction, being given both on the laptop and on paper.

Phase 3 During this phase the teams had 30 minutes to choose their three most effective and feasible concepts. These concepts could include multiple features or designs combined or developed into a single concept.

"I'd like to ask you to move onto the next stage. You now have 30 minutes to review your ideas and choose the best three."

Phase 4 The final phase of the experiment instructed the teams to draw and annotate their chosen ideas on single sheets of A3 paper, one idea per sheet. They were specifically told that the idea must be understood based on this piece of paper alone. This drawing technique was used to help streamline analysis and comparison of the ideas and to maintain the anonymity of the team members with respect to the expert assessor - removing possible marking bias.

"Please will you move onto the final stage and draw each idea onto a piece of paper in such a way that it is understandable without you having to be there to describe it."

The experiment controller was the same throughout the study and was instructed to behave neutrally, taking care to minimise experimental bias. Ideally the experiment controller would be hypothesis blind, however, due to pragmatic limitations this was not in this case. The controller spent the experiment reading on the corner of the Table 1 and was not allowed to interact with the participants unless to perform scripted actions. The five sessions were performed consecutively over two days. The participants were kept separate and incommunicado until after the end of the last session. Also the order in which the sessions were carried out was randomised.

3 Results and data handling

Each of the five experiments produced a single 4-channel video of the session in addition to the three A3 sheets with the final concepts, and any notes made by the participants. Notes at different phases were differentiated by changing the colour of the participant's pens at the start of each phase. This allowed the notes to be aligned with the video timeline and also let the researchers separate initial drawings or ideas from later additions.

The three metrics used to assess the teams performance were: total number of ideas, originality of ideas, and idea effectiveness. An idea count was generated from the audio discussion in the video and from the paper based sketches and notes. Care was taken to ensure that each idea was only counted once, as ideas were often discussed and then written down or recalled again later in the experiment. Idea originality was determined by comparing the teams final ideas and looking for similarities and differences. A strong commonality in ideas between all the teams would suggest few original ideas. An expert in eco-refrigerator design assessed idea effectiveness. Ideas that did not satisfy the brief were classed as irrelevant while ideas that the expert considered useful for reducing the wasted energy of inefficient product use were classed as effective. Since there are many ways of achieving the same function,

effective ideas using the same solution principle were grouped together to avoid counting multiple variants as totally different ideas.

Team	Total Number of Ideas Produced	Number (and Percentage) of Irrelevant Ideas		Number (and Percentage) of Effective Ideas	
Team 1 'Control'	94	22	(23%)	30	(32%)
Team 2 'Placebo'	47	10	(21%)	16	(34%)
Team 3 'Video'	40	5	(13%)	22	(55%)
Team 4 'Data'	39	2	(5%)	22	(56%)
Team 5 'Data + Clips'	57	3	(5%)	30	(53%)

Table 7. Idea comparison for the five teams

Table 7 summarises the results of these measures in percentage terms. Of the total number of ideas produced, the teams with relevant information (teams 3, 4 and 5) produced a higher proportion of more effective ideas than those without (55%, 56% and 53% compared to 32% and 34%). The teams with relevant information also had considerably fewer irrelevant ideas (13%, 5% and 5% compared to 23% and 21%).

In addition a qualitative review of how the teams referred to the additional information was carried out as part of an assessment of the relative usefulness of the information in the different formats. For example, the teams given video information never replayed the videos and seldom discussed what they had witnessed, whereas the teams with data information frequently returned to it, using it to prompt many discussions. The qualitative assessment centred on four aspects of the experiment and design process, with subsequent research questions for each. The aim of these questions was to shed additional light on the reasoning behind any conclusions from the quantitative work. These four aspects were:

1. Design Brief — Was the team's discussion of user behaviour and the design brief thorough and how did it affect their focus?
2. Idea Generation — How did the team perform with respect to idea generation and development?
3. Idea Evaluation — What approach did the teams use to assess and evaluate their ideas?
4. Input Information — Does the provided information appear to be useful to the team? How often do they refer to it? Does it steer the design process in anyway?

Although these results will be discussed in detail in a future publication it is important to outline them here in order to contextualise the discussion of the experimental method. One conclusion was that introducing additional information during ideation causes an interruption, negatively affecting the total

number of ideas generated. It was also found from the placebo that introducing irrelevant information, not only reduces the quantity, through interruption, but also showed no sign of improving the quality of ideas (team 2) – providing a baseline comparison for an interrupted ideation session. The no treatment control team (team 1), with no additional information, produced more ideas in total (97 compared to an average of 46) – providing a baseline for a standard ideation session. This initially appears to contradict creative stimuli literature (Goldschmidt & Tatsa, 2005) and support the premise that "quantity leads to quality" (Osborn, 1963; Reinig & Briggs, 2008). However, using both the no treatment baseline and placebo baseline it becomes apparent that had the other teams been given information in a way that was not an interruption they may also have gone on to produce as many ideas, but with the increased percentage of effective ideas that, arguably, additional information provides. Thus although the information may focus the team and reduce variety/quantity, the level of quality is increased which in this situation is beneficial. However, in less constrained tasks where pure variety is of paramount importance additional information could be detrimental.

In this experiment these two factors appear to balance out, with the no-treatment team producing as many effective ideas as teams 3, 4 and 5 simply through weight of numbers. Thus it may be prudent for future researches to attempt to introduce additional information in a less obtrusive manner, allow ideation to drop off before introduction, or introduce information at the outset. In this way it may be possible to gain the effectiveness benefits from the additional information without significantly reducing the total number of ideas produced.

4 Experimental review

In order to discuss the methodological successes or failings of this study it is first necessary to reflect on the problem areas being addressed: lack of contextualisation, system understanding, Method implementation, and control and normalisation (Table 2). This section examines each area separately to assess how well the study mitigated it and what additional techniques could have given further benefit. Finally this section brings together methodological findings from the study as a whole.

4.1 Contextualisation

With regard to context it was felt that the experiment performed adequately. Context was broken down into four areas: Social, cultural, activity and experimental. The participant population was contextualised with regard to its social makeup and culture. The population was selected to have a similar social structure to a work environment with the researchers taking the place of managers/meeting leaders. The participants were thus selected from a pool of relatively homogeneous experience, background, age and qualifications much as the majority of young employees in large companies. In this case the industrial and population contexts are relatively well matched with teams in both case being formed from larger groups of relatively homogeneous composition and limited

interpersonal relationships. This was established through a description of educational and social background as well as the use of the Belbin tests. Additional interpersonal relationship tests and background (sociometric and historical) could have revealed further information about possible problem parings or exceptional friends as discussed by Barrick et al. (1998), De Dreu & Weingart (2003) as well as providing a more detailed basis for comparison. However, this was deemed unnecessary as the researchers had personal knowledge of the participant population, their interpersonal relationships and backgrounds. This type of testing becomes more important when selecting from larger populations unknown to the researcher; where detailed information is required for statistical or qualitative comparisons. Non-homogeneous populations should be accounted for at the selection stage to ensure a representative participant group is produced. It is also worth noting that in small scale studies individuals will always vary and thus qualitative analysis can offer insights whereas statistical comparisons require larger populations to be effective.

Culturally there are differences between postgraduates and company employees; however, the participants were given a description of a hypothetical company structure and motivation. In this regard it is difficult to qualify the specific effect that this difference may have had on the results and as such is a clear limitation.

The context of the activity was given through the brief and the time and output pressures imposed on the teams. This could have been more detailed and achieved greater 'realism' through the use of some form of incentive or pressure to simulate the motivation/pressure of a company environment. This was balanced against the desire for a generalised task and time requirements. However, there is scope for improvement in this area particularly by relating it directly to specifically observed activity in industry. In this case, the selected activity (a design team brainstorming at the early stages) is accepted as common practice and as such relatively similar to most cases in industry.

Experimental context was recorded as part of the method planning and description. This covered the pre and post-test conditions, technology, methods and data handling procedures. This was an essential element in qualifying the significance of the results as well as allowing an informed judgment to be made about the value of the study. Recording this type of context is also critical for allowing the possibility of replication or reanalysis, an important requirement for community wide critique, validation and development. It should also be noted that this detailed contextualisation is an essential element for study replication and reliability.

4.2 Experimental system understanding

The main issue in this area was the difficulty associated with isolating individual variables at work in an experiment and from these establishing causal relationships. This study addressed this issue in a number

of ways. Multiple metrics (number, variety, and quality/effectiveness of ideas) were used to allow triangulation on the success criteria of 'benefit' here characterised by the overarching metric of team performance. In this way although each metric may not directly support one another (increased number does not support increased variation) they all support performance (increased number and increased variation both support increased performance). Thus triangulating metrics played a critical role in the assessment of team performance. Had only the metric 'number of ideas' been used, the conclusion could have been significantly different, namely, that additional information had a negative effect rather than the more balanced triangulated conclusion. Although this is not a substitute for true statistically significant study size, it does give more confidence in findings when multiple metrics agree in this fashion. In addition, comparing and contrasting these with the qualitative analysis gave detailed insights that would otherwise not have been possible from a small-scale study. The triangulation of the quantitative and qualitative analysis can also be used to assess the effectiveness of the control procedures – In this case, the qualitative analysis allowed the inertness of the placebo control to be established and also played an important role in assessing the usefulness of the no-treatment control. This was supplemented with intra-person reliability checks, carried out on the expert's evaluation of idea relevance and effectiveness – the assessor remarked a random selection of the ideas to ensure that they were consistent from start to finish.

The placebo development process entailed a systematic consideration and classification of the underlying variables and allowed further refinement of system understanding (see Section 4.4). Finally, a deviant case analysis was carried out focusing on identifying evidence contrary to the experimental hypothesis and then attempting to explain this. This can prove to be a powerful technique for revealing conclusion fallacies and other experimental problems. In this case the deviant case was the large number of ideas produced by the no-treatment control team. In attempting to explain this issue a great deal of insight into the wider implications of the study intervention was generated, most importantly the impact of interruption. This also demonstrates the power of triangulating multiple metrics and techniques; although the quantity metric deviated from the hypothesis both of the other quantitative metrics and the qualitative analysis supported the hypothesis. The additional assessment of what caused this discrepancy through the deviant case analysis also supported the final hypothesis by reassessing exactly what variables affected idea count.

This area could be improved by using additional methods and metrics, such as post-test interviews to examine participant experience or the perceived usefulness of the information. However, as additional methods and metrics are introduced, demands on analysis time increase rapidly.

4.3 Method implementation

The main issue with method implementation in design research was the development and use of inadequately defined, non-comparable idiosyncratic research methods through a failure to define populations, terms, techniques and lack of standardisation. This study addresses these by defining both the overall methods in this paper and the specific control methods in online. In addition care was taken to thoroughly define context, terms, techniques and environment throughout the study. Standardisation was promoted by the use of commonly available techniques such as the Belbin team roles and the provision of any new techniques developed for this study. This makes use of methods freely available and previously validated (Belbin) and in cases where this is not possible defines methods in sufficient detail as to allow them to be replicated or validated by a third party. The methods and supporting materials such as scripts are freely available on the website: www.designresearchmethods.com.

4.4 Controls and normalisation

The main issue in this area was the lack of effective normalisation for experimental effects caused by insufficient or inappropriate use of control groups. In order to address this, the study introduced a placebo control group in addition to the no-treatment control group more commonly used in design research. This had several advantages over the standard no-treatment control used alone. Firstly, the placebo allowed for the normalisation and removal of experimental effects other than those directly under study such as interruption. Secondly, the two control baselines used in conjunction allowed the affect under study to be isolated effectively from other experimental factors such as interruption.

It should be noted that the effectiveness of the placebo control group was assessed qualitatively before it was accepted for use as a baseline for performance. The qualitative assessment examined several key areas that could have rendered the placebo ineffective. It was seen that the placebo video did not engender any suspicion or unusual dialog amongst the participants and was also watched with a similar level of attentiveness as the other video teams causing a similar level of disruption. Also, despite the video being watched attentively, the placebo was not referred back to during the later stages of the study compared to the other videos. This implies that the video did indeed contain no obviously relevant information for the design task. Taking this assessment into account it was felt that the placebo did indeed provide a valid baseline against which to compare the other teams.

The qualitative review of the teams' performance was based on the assessment of four topics: discussion of user behaviour and the brief; idea generation and development; idea assessment and evaluation; and input information review and discussion. These topics provided an overview of the teams understanding of the brief, their use of creativity tools and approaches, how they reviewed their ideas and how they interacted with the input information.

The control teams (1 and 2) began by rapidly developing ideas while the treatment teams (3, 4 and 5) were slower in the first phase (Table 8). This was mainly caused by a more detailed review of the brief and attempts to draw up lists of possible causes of bad behaviours. This was particularly evident in team 4, producing only 3 ideas in phase 1, who deliberately stopped all ideation until they were satisfied with their review of the problem, cutting short members who deviated from this goal. It is, therefore, difficult to quantitatively compare the teams in phases 1 and 2, as the approaches taken in phase 1 by three of the teams specifically limited ideation.

However, in phase 2 all five teams increased their rate of idea generation, producing the main bulk of the total ideas created. Comparing the teams at this time showed that, once ideating, all the teams were relatively similar in the rate at which they produced ideas with the exception of the placebo control team. It is also important to note that although the total number of ideas produced at this phase is similar, the percentage of those deemed relevant and high quality were significantly greater in the treatment teams (Table 7). It is also important to note the significant increase in the percentage of effective ideas after the introduction of the stimuli compared to the more uniform profile of effective ideas shown by the placebo and no-treatment teams.

Other interesting points of note from this qualitative assessment are: team 4's deliberate limitation of creativity in phase 1; teams' 2 and 3 sat in silence while watching of the videos, recording no ideas either on paper or verbally – detrimentally affecting their number of ideas; the continued high level of idea generation by team 1 in phases 3 and 4, and the dominant role of the shaper in teams 1 and 4, pushing team 1 forward every few minutes or strictly controlling team 4's ideation respectively. It is also interesting to note that team 1 appeared to be having the most 'fun', deliberately not limiting themselves on feasibility grounds – this possibly explains, to some extent, the high number and long duration of team 1's ideation.

	Total ideas produced / percentage of effective ideas			**Total Number of Ideas Produced**
Team	**Phase 1**	**Phase 2**	**Phase 3 & 4**	
Team 1 'Control'	21 / 43%	42 / 29%	31 / 29%	94
Team 2 'Placebo'	20 / 35%	22 / 36%	6 / 17%	47
Team 3 'Video'	10 / 30%	22 / 77%	8 / 25%	40
Team 4 'Data'	3 / 33%	28 / 71%	8 / 13%	39
Team 5 'Data + Clips'	9 / 78%	32 / 50%	16 / 44%	57

Table 8. Idea v. phase comparison for the five teams

This assessment also brought to light the fact that teams 2 and 3 referred back to the video only once or twice, whilst teams 4 and 5, often referred to the data. This suggests a level of fixation with the data; although excessive fixation can negatively effect ideation – reducing creativity – some fixation on the solution can benefit the teams – focusing their efforts – and in this case provide a reason for their high percentage of effective ideas. Also, the lack of solution specific information caused the two control teams to generate a greater percentage of irrelevant and ineffective ideas, reinforcing the value of the placebo team as an effective form of control.

4.5 Study overview

In order to assess the methodological quality of this study the authors have taken four critical routes. Firstly, the study is compared to other study types and the tradeoffs are discussed. Secondly, the study is discussed with respect to an independently generated set of metrics for assessing study quality. Thirdly, the study is compared to two closely related studies. Finally, this is summarised with respect to the problems outlined in Table 2.

Comparison against study types

The first route to assessing the quality and limitations of the methods reported here is comparing them to the other types of study. This study falls at one end of a spectrum ranging from large statistically significant studies to single person case studies. Across this range are a number of tradeoffs, most notably in the types of insight that can be elucidated about causal relationships and external validity. Table 9 summarises the broad attributes of large, medium and small-scale studies. Although this study clearly falls into the ‘small’ category in Table 9 careful research design can strengthen many of the types of validity, reliability and replicability. In this case the use of multiple metrics and control groups has allowed for a better distinction between opinion and results improving the conclusion validity.

There is also a spectrum across studies in terms of contrivance, varying from fully embedded ethnographic type work to highly contrived laboratory studies. Again, there are tradeoffs across this range, most critically in the level of external validity or reliability and the level of internal validity and replicability. This study was highly contrived with the participants being limited to the resources given in the room. It was felt that this was appropriate as typical ideation sessions of this sort were effectively ‘cut-off’ not using phones or external resources, and also taking place over a limited time period in a predefined room with set goals. Thus, although in some cases this could be a detrimental trade-off it was considered to be appropriate in this case considering issues of reliability (Cross, et al., 1996). Despite this, reliability has been improved through selection of a population similar to industry and the description of context (Section 4.1) allowing some conclusions to be made regarding its relation to industrial scenarios. It should be noted that methodological rigour is not only critical to small-scale

studies but to all empirical studies and the points highlighted throughout this paper apply across the range of size and contrivance outlined in Table 9.

Size	**Small**	**Medium**	**Large**
Description	A study with too few data points to effectively statistics e.g. (Cai, et al., 2010)	A study using sufficient data points to allow non-parametric statistics e.g. (Magin & Churches, 1995)	A study with a high number of data points allowing parametric statistical e.g. (McCarney, et al., 2007)
Types of Validity			
Internal	Can give non-statistical insight into causal relationships	Can statistically identify causal relationships in a limited population	Can statistically identify and quantify causal relationships
Causal construct	Can offer insight but can not offer measures	Can offer insight and can offer limited measures	Can offer insight and can offer explicit measures
Statistical	N/A	Non-parametric	Parametric
External	Can give non-statistical insights which can inform wider understanding	Can not be generalised outside the sample population	Can be generalised across populations using statistical models
Conclusion	Difficult to differentiate opinion from results	Clearer split between results and opinion	Clear split between results and opinion
Replicability	Difficult to replicate as results are highly dependant on participants	Can be replicated but is dependant on population	Can be replicated as long as population selection is consistent
Reliability	Usually only applies to the specific context of the study	Can apply to a wider context but still limited	Applies to the whole population being modelled
Pragmatic considerations	Small size can make setup and capture easier however analysis needs further interpretation	Medium size demands moderate setup but can make data analysis simpler in regard to data set size	Large size demands extensive setup and can make analysis complex due to the size of the data set

Table 9: A comparison of study types

Comparison against established methodological metrics

The second route to assessing the quality of the research methods reported here is comparing them to existing measures in the literature. Dyba and Dingsoyr (2008) outline 11 such metrics in their assessment of empirical studies in software development. Of these, 6 relate to Method implementation (the others relate to reporting and contribution). This analysis of method was also heavily informed by Klein and Myers' (1999) 'principals for interpretive field research' which also emphasise elements such as contextualisation, researcher interaction, generalisation and bias. Table 10 summarises each of these measures and how they were addressed in this study.

Measures (Quoted from Dyba and Dingsoyr p.839)	**How it was addressed**
There was an adequate description for the context in which the research was carried out	An attempt was made to describe the social, cultural and activity context of the study – Section 2. The environment, methods and population are explicitly described throughout this paper.
There was adequate description of the sample used and the methods for identifying and recruiting the sample	Participants were selected randomly from a described population, sorted into teams semi-randomly using Belbin and other metrics and allocated treatments randomly – Section 2.1
Any control groups were used to compare treatments	Both no-treatment and placebo groups used as baselines and compared against the other teams – Section 4.4
Appropriate data collection methods were used and described	The data collection methods are explicitly stated with diagrams and an explanation of setup – Section 2
There was adequate description of the methods used to analyse the data and whether appropriate methods for ensuring the data analysis were grounded in the data	The experimental procedure – Section 2.2 – and analysis methods were described explicitly. Quantitative results are presented alongside qualitative discussions – Section 3
The relationship between the researcher and participants was considered to an adequate degree	The interactions between the researcher and the participant were tightly controlled and in most cases explicit scripts for interactions was provided – Section 2.2

Table 10: A critique using Dyba and Dingsoyr's (2008) metrics for assessing empirical studies

Assessing these against the detailed criteria/checklist presented by Dyba and Dingsoyr (adapted for design) this study rates positively for all the methodological metrics. In addition to the table the assessment highlighted two, already identified, shortcomings: a) the sample size was too small to allow statistical analysis; b) the difficulty in accessing the differences between the control and treatment teams. In addition, the study could have benefited from a more rigorous approach to assessing and matching context relative to industry. This formed a key part of the discussion at the Delft protocols workshops (Cross, et al., 1996) and is a key issue in design research in general. However, the study made explicit the nature of the activity, carefully selected a population with a similar social structure to that in industry and attempted to elucidate the cultural context of the 'company' using the brief. On reflection, it is clear that these could have been improved by selecting a specific population for comparison (e.g. a specific company).

Comparison against related studies

The final route to assessing the quality of the methods reported here is comparing this study to closely related studies. The importance of the placebo for isolating key factors is highlighted by the work of Collado-Ruiz and Ostad-Ahmad-Ghorabi (2010) – not using a placebo control. Both studies initially

found that the introduction of additional stimuli had a detrimental effect on the number of ideas generated (with the no-treatment teams outperforming the test teams). This supports this study as these agree with the findings when not using a placebo control group. However, the placebo and additional metrics, used in the study reported here, allowed for the identification and normalisation of the interruption effect on team ideation. This subsequently demonstrated that additional information adversely effects idea quantity but actually improves idea quality and effectiveness. The important role of the control groups is also highlighted by looking at Lopez-Mesa et al. (2009) who do not use them at all. Lopez-Mesa et al. also support the quality findings of this study; showing that visual stimuli give improved quality and variety of ideas compared to general questions.

Comparison against the established problems

Finally, Table 11 summarises the mitigation approaches discussed in this section and also highlights some of the key limitations. One such limitation was the influence of the experiment controller. Despite being scripted and accounted for through the use the placebo control, there were several deficiencies. Firstly, the Q and A session was not scripted due to a limited test development period. Secondly, the controller was not hypothesis blind – where the participant or researcher are kept ignorant of the experimental hypothesis both before and during an experiment (Adelman, 1991). Finally, there was no analysis of whether the participants remained hypothesis blind post-test, although they were hypothesis blind pre-test.

Problem	Mitigation	Limitations
Lack of Context	Social context described, activity context provided, cultural context described, experimental setup/method described,	Description and control of context could be more sophisticated, specific populations could be specified
System Understanding	Using a systematic method for classifying the variables, triangulating multiple metrics both qualitative and quantitative	Additional metrics could have been added, reliability affected by sample size, general scope limits specificity
Method implementation	Full disclosure of methods for the study and the placebo control, standardisation, triangulation, thorough critique of methods	More time could have been spent on prototyping allowing more flexible interactions between research and subject
Control and Normalisation	The use of both no-treatment and placebo controls, detailed deviant cases analysis, strict control of researcher interaction	Independent validation of the placebo control in this context could have been beneficial

Table 11: Summary of mitigating techniques and limitations

5.0 Conclusions

In conclusion, this paper highlights the usefulness of small-scale studies when conducted using rigorous methods. Relating back to the papers goals – demonstration of approaches and their benefits – the study

benefited significantly from the use of control procedures (particularly the placebo), triangulation of metrics (qualitative and quantitative) and detailed self-critique. Through these approaches it was possible to improve validity without significant additional experimentation. The benefits of these approaches, normally underutilised in design research, have been revealed through a detailed assessment against existing study types, established methodological metrics and analogous studies. This has emphasised that although small-scale studies are not a substitute for larger statistical studies there is a clear use for them in identifying trends and possible research directions.

This work has attempted to cohesively address context, system understanding, Method implementation, and control and normalisation. Triangulating qualitative and quantitative data in addition to improved controls not only allowed improved validation of the test hypothesis but also gave a measure of validity to the control techniques used. Although there are areas for further improvement – particularly the independent validation of the placebo development method – these techniques offer improvement over conventional studies of this type. Techniques include: triangulation of metrics, a placebo control team and detailed deviant case analysis and critique. A breakdown of these methods has been discussed in this paper and made available online in order to promote replication, standardisation and shared understanding.

The secondary goal of this paper was to identify and specify further work that would allow a more rigorous link to be drawn between small-scale design experimental research and industrial practice. It is ultimately not possible to fully define the reliability of the study when considering the findings in relation to an industrial context. Thus we propose a larger study involving the direct comparison of analogous situations across levels of contrivance. To this end, the authors identify three studies – an ethnographic study (capturing design situations in industry) – the replication of multiple design situations in a contrived laboratory setting – and finally the replication of these contrived design situations using engineers in an industrial setting. In this way, three levels of contrivance can be related systematically and linked providing a basis for comparison between small-scale empirical design research and design practice in industry (Cash, et al., 2011).

Acknowledgements

The work reported in this paper has been undertaken as part of the EPSRC Innovative Manufacturing Research Centre at the University of Bath (grant reference GR/R67507/0) and has been supported by a number of industrial companies. The authors gratefully acknowledge this support and express their thanks for the advice and support of all concerned. The authors would also like to acknowledge the extremely thorough and valuable feedback provided by the reviewers, which played a vital role in make this work, what it is now.

References

Adair, J. G., Sharpe, D., & Huynh, C. L. (1990). The placebo control group: An analysis of its effectiveness in educational research. *The Journal of Experimental Education, 59*, 67-86.

Adelman, L. (1991). Experiments, quasi-experiments, and case studies: a review of empirical methods for evaluating decision support systems. *IEEE transactions on systems, man, and cybernetics, 21*, 293-301.

Ancona, D. G., & Caldwell, D. F. (1992). Demography and design: Predictors of new product team performance. *Organization Science, 3*, 321-341.

Bakeman, R., & Deckner, D. F. (2003). Analysis of behaviour streams. In D. M. Teti (Ed.), *Handbook of Research Methods in Developmental Psychology*. Oxford, U.K.: Blackwell.

Ball, L. J., & Ormerod, T. C. (2000). Applying ethnography in the analysis and support of expertise in engineering design. *Design Studies, 21*, 403-421.

Baltes, B. B., Dickson, M. W., Sherman, M. P., Bauer, C. C., & LaGanke, J. S. (2002). Computer-Mediated Communication and Group Decision Making: A Meta-Analysis* 1. *Organizational behavior and human decision processes, 87*, 156-179.

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of applied psychology, 83*, 377-391.

Belbin. (2010). Belbin homepage [Online]. In: www.belbin.com.

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of applied psychology, 92*, 595-615.

Bender, B., Reinicke, T., Wunsche, T., & Blessing, L. T. M. (2002). Application of methods from social sciences in design research. In *Design 2002 International Design Conference*. Dubrovnik, Croatia.

Blessing, L. T. M., & Chakrabarti, A. (2009). *DRM, a Design Research Methodology*. New York: Springer.

Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of personality and social psychology, 50*, 543-549.

Briggs, R. O. (2006). On theory-driven design and deployment of collaboration systems. *International Journal of Human-Computer Studies, 64*, 573-582.

Cai, H., Do, E. Y. L., & Zimring, C. M. (2010). Extended linkography and distance graph in design evaluation: an empirical study of the dual effects of inspiration sources in creative design. *Design Studies, 31*, 146-168.

Campion, M. A. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel psychology, 46*, 823-850.

Cash, P., Hicks, B. J., & Culley, S. J. (2009). The challenges facing ethnographic design research: A proposed methodological solution. In *ICED 09 International Conference on Engineering Design*. Stanford, CA, USA.

Cash, P., Hicks, B. J., Culley, S. J., & Salustri, F. (2011). Designer behaviour and activity: An industrial observation method. In *ICED 11 International conference on engineering design*. Copenhagen, Denmark.

Collado-Ruiz, D., & Ostad-Ahmad-Ghorabi, H. (2010). Influence of environmental information on creativity. *Design Studies, 31*, 479-498.

Corremans, J. A. M. (2009). Measuring the effectiveness of a design method to generate form alternatives: an experiment performed with freshmen students product development. *Journal of Engineering Design, 22*, 259-274.

Cross, N. (2001). Design cognition: Results from protocol and other empirical studies of design activity. In C. Eastman, W. Newstetter & M. McCracken (Eds.), *Design knowing and learning: Cognition in design education* (pp. 79-103). Amsterdam: Elsevier science.

Cross, N. (2007). Forty years of design research. *Design Studies, 28*, 1-4.

Cross, N., Christiaans, H., & Dorst, K. (1996). *Analysing design activity*. Chichester: John Wiley and Sons, UK.

Cummings, L. L., Huber, G. P., & Arendt, E. (1974). Effects of size and spatial arrangements on group decision making. *The Academy of Management Journal, 17*, 460-475.

D'Astous, P., Robillard, P. N., Detienne, F., & Visser, W. (2001). Quantitative measurements of the influence of participant roles during peer review meetings. *Empirical software engineering, 6*, 143-159.

De Dreu, C. K. W., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of applied psychology, 88*, 741-749.

Dillon, P. (2006). Creativity, integrativism and a pedagogy of connection. *Thinking Skills and Creativity, 1*, 69-83.

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies, 26*, 445-461.

Drach-Zahavy, A., & Anit, S. (2001). Understanding team innovation: The role of team processes and structures. *Group dynamics, 5*, 111-123.

Dyba, T., & Dingsoyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology, 50*, 833-859.

Elias, E. W. A., Dekoninck, E., & Culley, S. J. (2009). Designing for ˜"use phase" energy losses of domestic products. *Proceedings of the Institution of Mechanical Engineers; Part B; Journal of Engineering Manufacture, 223*, 115-122.

Friedman, K. (2003). Theory construction in design research: Criteria: Approaches, and methods. *Design Studies, 24*, 507-522.

Gero, J. S., & Tang, H.-H. (2001). The differences between retrospective and concurrent protocols in revealing the process-oriented aspects of the design process. *Design Studies, 22*, 283-295.

Glasgow, R. E., & Emmons, K. M. (2007). How can we increase translation of research into practice? Types of evidence needed. *Annual Review of Public Health, 28*, 413-433.

Goldschmidt, G., & Tatsa, D. (2005). How good are good ideas? Correlates of design creativity. *Design Studies, 26*, 593-611.

Gorard, S., & Cook, T. D. (2007). Where does good evidence come from? *International Journal of Research & Method in Education, 30*, 307-323.

Gorla, N., & Lam, Y. W. (2004). Who should work with whom?: Building effective software project teams. *Communications of the ACM, 47*, 79-82.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-computer interaction, 13*, 203-261.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology, 47*, 307-338.

Hackman, J. R., & Vidmar, N. (1970). Effects of size and task type on group performance and member reactions. *Sociometry, 33*, 37-54.

Hansen, P. K., Mabogunjie, A., Eris, O., & Leifer, L. (2001). The product development process ontology: Creating a learning research community. In *ICED 01 International Conference on Engineering Design*. Glasgow, UK.

Hare, A. P. (1952). A study of interaction and consensus in different sized groups. *American Sociological Review*, 261-267.

Henry, S. M., & Stevens, K. T. (1999). Using Belbin's leadership role to improve team effectiveness: An empirical investigation. *The Journal of systems and software, 44*, 241-250.

Hoffman, L. R. (1965). Group problem solving. In *Advances in experimental social psychology* (Vol. 2, pp. 99-131). New York: Academic press.

Howard, T. J., Culley, S. J., & Dekoninck, E. (2010). Reuse of ideas and concepts for creative stimuli in engineering design. *Journal of Engineering Design, 0*, 1-17.

Hwang, H. G., & Guynes, J. (1994). The effect of group size on group performance in computer-supported decision making. *Information & management, 26*, 189-198.

Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El-Emam, K., & Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering, 28*, 721-734.

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, 67-93.

Kurtoglu, T., Campbell, M. I., & Linsey, J. S. (2009). An experimental study on the effects of a computational design tool on concept generation. *Design Studies, 30*, 676-703.

Lanubile, F. (1997). Empirical evaluation of software maintenance technologies. *Empirical software engineering, 2*, 97-108.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life*. New York: Cambridge University Press.

Leber, P. (2000). The use of placebo control groups in the assessment of psychiatric drugs: an historical context. *Biological Psychiatry, 47*, 699-706.

Lemons, G., Carberry, A., Swan, C., Jarvin, L., & Rogers, C. (2010). The benefits of model building in teaching engineering design. *Design Studies, 31*, 288-309.

Lopez-Mesa, B., Mulet, E., Vidal, R., & Thompson, G. (2009). Effects of additional stimuli on idea-finding in design teams. *Journal of Engineering Design, 22*, 31-54.

Magin, D. J., & Churches, A. E. (1995). Peer tutoring in engineering design: A case study. *Studies in Higher Education, 20*, 73-85.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC medical research methodology, 7*, 30-37.

Onwuegbuzie, A. J., & Leech, N. L. (2006). Linking research questions to mixed methods data analysis procedures. *The Qualitative report, 11*, 474-498.

Osborn, A. F. (1963). *Applied imagination: Principles and procedures of creative problem-solving*. New York: Scribner.

Reinig, B. A., & Briggs, R. O. (2008). On the relationship between idea-quantity and idea-quality during ideation. *Group Decision and Negotiation, 17*, 403-420.

Robinson, H., Segal, J., & Sharp, H. (2007). Ethnographically-informed empirical studies of software practice. *Information and Software Technology, 49*, 540-551.

Robinson, M. A. (2010). Work sampling: Methodological advances and new applications. *Human Factors and Ergonomics in Manufacturing & Service Industries, 20*, 42-60.

Robinson, M. A., Sparrow, P. R., Clegg, C., & Birdi, K. (2005). Design engineering competencies: Future requirements and predicted changes in the forthcoming decade. *Design Studies, 26*, 123-153.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors, 50*, 903-933.

Senior, B. (1997). Team roles and team performance: Is there 'really' a link? *Journal of Occupational and Organizational Psychology, 70*, 241-258.

Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies, 24*, 111-134.

Sharp, H., & Robinson, H. (2008). Collaboration and co-ordination in mature eXtreme programming teams. *International Journal of Human-Computer Studies, 66*, 506-518.

Stempfle, J., & Badke-schaub, P. (2002). Thinking in design teams-an analysis of team communication. *Design Studies, 23*, 473-496.

Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management, 32*, 29.

Stones, C., & Cassidy, T. (2010). Seeing and discovering: how do student designers reinterpret sketches and digital marks during graphic design ideation? *Design Studies, 31*, 439-460.

Torgerson, D. J., & Torgerson, C. J. (2003). Avoiding bias in randomised controlled trials in educational research. *British journal of educational studies, 51*, 36-45.

Valkenburg, R., & Kleinsmann, M. (2009). Performing high quality research into design practice. In *ICED'09 International conference on engineering design*. Stanford, CA, USA.